Recuperación de Información y Recomendaciones en la Web 2020

Docente

Libertad Tansini

Autores

Alejandro Flocken 3.440.329-6

Jerónimo Florit 4.333.114-5

Introducción

En el marco del desarrollo de una tarea que aplique los conocimientos y técnicas tratados en la materia, nos planteamos el objetivo de recuperar información relativa al accionar legislativo de parlamentarios y otros gobernantes.

Identificamos como un tema de interés ciudadano el poder contar con información procesada, respecto al nivel de producción en materia legislativa, que permita realizar distintos tipos de análisis de forma simplificada.

Como exploración inicial decidimos relevar la viabilidad de usar como fuentes de información los sitios web del propio Parlamento, así como de la Dirección Nacional de Impresiones y Publicaciones Oficiales (IMPO). Encontramos que el IMPO brinda el acceso como Datos Abiertos a toda la información pública disponible en sus bases de datos, a través de una API Rest que nos permitirá obtener información de forma muy conveniente, por lo cual decidimos avanzar utilizando esa fuente (https://www.impo.com.uy).

Si bien en dicho sitio se publica la información de leyes promulgadas, no disponibiliza mecanismos para realizar el seguimiento de la actividad parlamentaria de un determinado parlamentario ni obtener estadísticas sobre su participación.

Por lo tanto, consideramos interesante realizar una implementación que permita analizar participaciones, realizar búsquedas avanzadas, evaluar asociaciones entre actores y/o temáticas, generar información estadística y en general hacer un uso exhaustivo y novedoso de la información disponible.

Análisis inicial de la solución

Del relevamiento realizado surgió la posibilidad de usar la API REST provista por el IMPO [1], que disponibiliza la información en formato JSON.

El esquema definido por IMPO se encuentra disponible en la dirección https://www.impo.com.uy/resources/basesIMPO.json y contiene atributos como por ejemplo:

```
"tipoNorma": {
        "type": "string",
        "description" : "Tipo de norma o aviso"
    },
"nroNorma": {
        "type": "string",
        "description" : "Número de la norma o aviso"
    },
```

Y particularmente el siguiente que es de interés para nuestra implementación:

```
"firmantes": {
        "type": "string",
"description" : "Firmantes"
```

Extracción de la información

Como parte de la solución proponemos implementar un Demonio que realice consultas diariamente, recorriendo números de las diferentes leyes promulgadas, utilizando el formato disponible:

https://www.impo.com.uy/bases/leyes/{{NroLey}}-{{año}}?json=true

En un alcance inicial de la tarea, proponemos acotar el estudio a las normas promulgadas durante 2019 y 2020, dejando abierta la posibilidad futura de integrar las de todos los años.

La información recolectada será parseada y tratada para almacenar en una BD que crearemos, indexando la información de forma conveniente para luego poder consultarla y analizarla con el tipo de cruzamiento que queremos lograr.

Mediante una web vamos a ofrecer información estadística sobre la actuación parlamentaria de determinado legislador (cantidad de leyes firmadas, conjuntamente a quienes, cantidad de artículos por ley, etc), también permitiendo búsquedas por tema, firmante, etc.

[1] https://www.impo.com.uy/datos-abiertos/

Arquitectura de la solución

A nivel de la recuperación y explotación de la información, la solución propuesta cuenta con 4 actividades básicas en relación a los datos:

- Extracción
- Transformación
- Carga
- Consulta

Extracción

La forma implementada para la recuperación de la información está basada en las características de los datos a procesar, en este caso el cuerpo de leyes publicadas en años anteriores es estático, en un conjunto que no cambia y que se puede extraer en un único proceso inicial. Luego el cuerpo debe ir creciendo con las nuevas leyes publicadas.

Dadas estas características, se modificó la visión inicial de establecer un demonio *polleando* por información, a un enfoque de un proceso base que extraiga toda la información histórica, que pueda luego ser complementado con una ejecución en régimen diario, para agregar la nueva información disponible en el año en curso.

Transformación

El proceso de extracción nos entrega información en formato JSON, que si bien las BD permiten almacenar directamente en esa forma, a efectos del tratamiento que vamos a darle a estos datos, debe ser transformada a las distintas entidades con las que se trabajará.

Por lo tanto, previo a la carga se realiza un proceso *json_decode*, complementado con transformación de la codificación a ISO-8859-1, que nos permite visualizar y persistir correctamente información en idioma español.

Carga

Para la tarea de persistir la información se optó por una BD relacional MySQL, por ser una opción estable, de acceso libre y bien conocida.

De los datos recuperados y transformados, se identificaron 3 entidades fuertes a representar, junto a sus relaciones:

- Normas: Son las leyes que se registrarán en el sistema, contienen entre otros los siguientes atributos:
    - Tipo: En este caso siempre será 1 - Ley
    - Número de norma: La identifican únicamente
    - Año: Toda la normativa está organizada agrupada en años
    - Nombre: El nombre de la ley describe brevemente su contenido

- Artículos: Son las partes que componen cada ley, siempre existe al menos uno. Tienen, entre otros, los siguientes atributos:
    - nro Artículo: Es la disposición del artículo dentro de la ley

    - texto Artículo: Es el contenido del artículo, todos juntos conforman el texto de la ley.
    - notas, referencias, link

- Firmante: Esta entidad es de particular interés para nosotros y no es retornada de una forma estructurada, sino como una cadena de texto que contiene múltiples personas. Por lo que para poblar esta tabla tuvimos que hacer un procesamiento de la información recibida. La entidad registra únicamente el nombre del firmante. Tan relevante como el registro del firmante, es registrar la asociación de estos con las distintas normas, para eso se creó la tabla **normafirmante** que vincula las 2 entidades. Esta construcción sería equivalente a un índice invertido, dónde el término es el firmante y el documento la norma referida. Esta construcción nos permitirá más adelante hacer el análisis y recuperación de información por firmante, de forma eficiente.

Además de los datos recuperados directamente se cargaron otras entidades necesarias para la solución

- Tipo: Es el tipo de norma referido anteriormente, en el alcance de esta solución solo se trabaja con el tipo 1-Ley, pero en un trabajo futuro podrían alcanzarse otros tipos de normativas.
- Publicaciones: Tabla que contiene para cada año pasado, la mínima y máxima norma de ese año, permite optimizar la extracción inicial de datos.
- Parámetros: Permite controlar los años sobre los que se quiere trabajar la extracción

Consulta

La consulta se ofrece a los usuarios a través de una web simple, que presenta inicialmente 4 funcionalidades que se describirán en más detalle más adelante:

- Consulta de normas por Año y Número
- Evaluación estadística de la producción de los distintos firmantes
- Búsqueda temática dentro de las normas
- Normativas por firmante

Tecnologías utilizadas

Se trabajó con las siguientes tecnologías:

- XAMPP
  - Apache
  - MySQL
  - PHP
- Otras tecnologías para el desarrollo web
  - HTML
  - Javascript
  - Ajax
  - CSS
  - Twig
- Entorno de desarrollo
  - Notepad++
  - Chrome

Funcionalidades Provistas

Recuperación de la información

Se logró recuperar la información de todas las leyes publicadas entre 2005 y 2020, registrando todas las entidades identificadas anteriormente. La estrategia de recuperación se apoyó en la construcción de una tabla con topes por año llamada publicaciones y cargada manualmente a partir de la consulta de estos datos al IMPO. Esta tabla le marca al proceso el rango de normas que debe recorrer por año. Dado que a priori no se conocen las normas publicadas por año, el proceso debe recorrer secuencialmente una lista, registrando aquellas para las que la API retorna datos.

En la imagen de la derecha se visualizan resultados cuantitativos de la recuperación de la información.

anio	norma_desde	norma_hasta
2005	17862	17938
2006	17938	18084
2007	18084	18247
2008	18247	18456
2009	18456	18639
2010	18639	18726
2011	18726	18881
2012	18881	19044
2013	19044	19186
2014	19186	19307
2015	19307	19367
2016	19367	19474
2017	19474	19591
2018	19591	19733
2019	19733	19861
2020	19861	19913

count(*)	norma
1412	norma
5595	articulo
112	firmante
8035	norma_firmante
1254	referencias

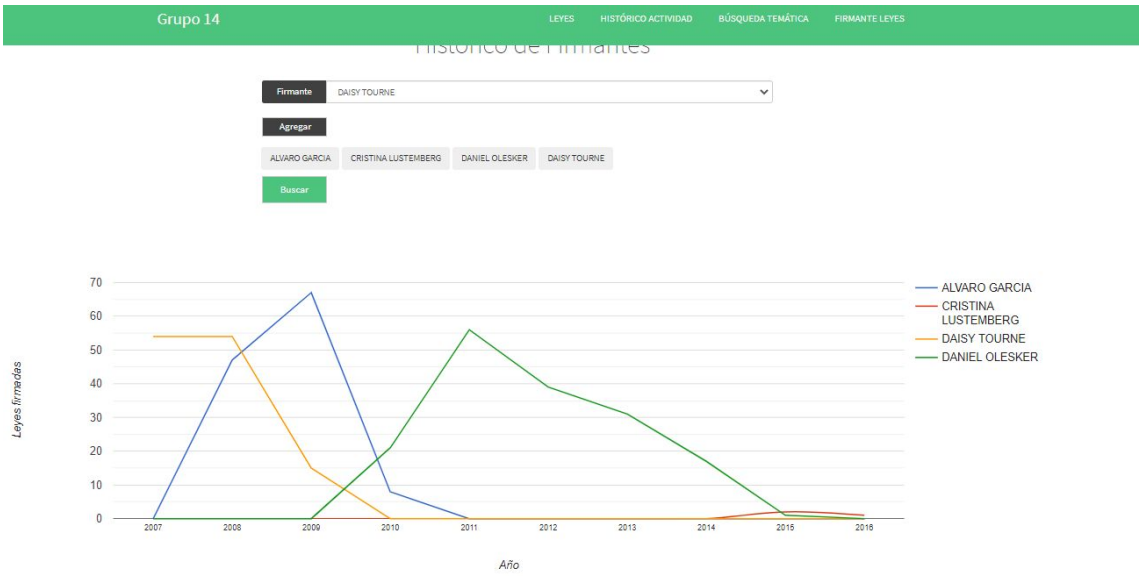
Consulta de Normas

En un primer nivel de uso de la plataforma, se construyó una búsqueda explícita de normas, por años de publicación y/o número de norma. Esta búsqueda permite al usuario una primera aproximación a la norma que está buscando, si cuenta con el dato preciso de número, o si conoce o le interesa saber las publicaciones de determinado año. El resultado permite visualizar Número de norma, Año, Nombre, Fecha de publicación y también enlazar directamente con el sitio del IMPO que contiene toda la norma.

Histórico de producción por firmantes

En esta funcionalidad ofrecimos al usuario la posibilidad de analizar gráficamente y a través de tablas de datos la producción comparativa entre los distintos firmantes.

La pantalla presenta un cuadro de selección dónde se pueden elegir múltiples firmantes de la lista completa que el sistema tiene registrado, para luego comparar sus producciones a lo largo del tiempo.



Firmante	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	Maximo	Minimo	Promedio
DAISY TOURNE	54	54	15	0	0	0	0	0	0	0	54	0	12.3
ALVARO GARCIA	0	47	67	8	0	0	0	0	0	0	67	0	12.2
CRISTINA LUSTEMBERG	0	0	0	0	0	0	0	0	2	1	2	0	0.3
DANIEL OLESKER	0	0	0	21	56	39	31	17	1	0	56	0	16.5

Búsqueda temática

Se generó una pantalla de búsqueda temática, permitiendo aplicar filtros de año y/o número de norma, pero especialmente incorporando la búsqueda de términos dentro de las normativas disponibles.

La funcionalidad permite la búsqueda con conceptos, como por ejemplo “Poder Ejecutivo”, así como de conjuntos de términos y/o conceptos que deben aparecer en la normativa buscada. Para lograr esto se consideró el punto y coma como carácter separador, con lo cual un usuario podría buscar artículos dentro de las distintas normativas que contengan por ejemplo el concepto “partidos políticos” y también las palabras “sexos” y “equitativa” ingresando la búsqueda ***partidos políticos;sexos;equitativa***

Producción legislativa de un firmante

La última funcionalidad presentada permite visualizar completamente la producción legislativa de un determinado firmante, que se puede seleccionar de un combo con todos los firmantes registrados en el sistema.

Funcionalidades en desarrollo

Índice invertido

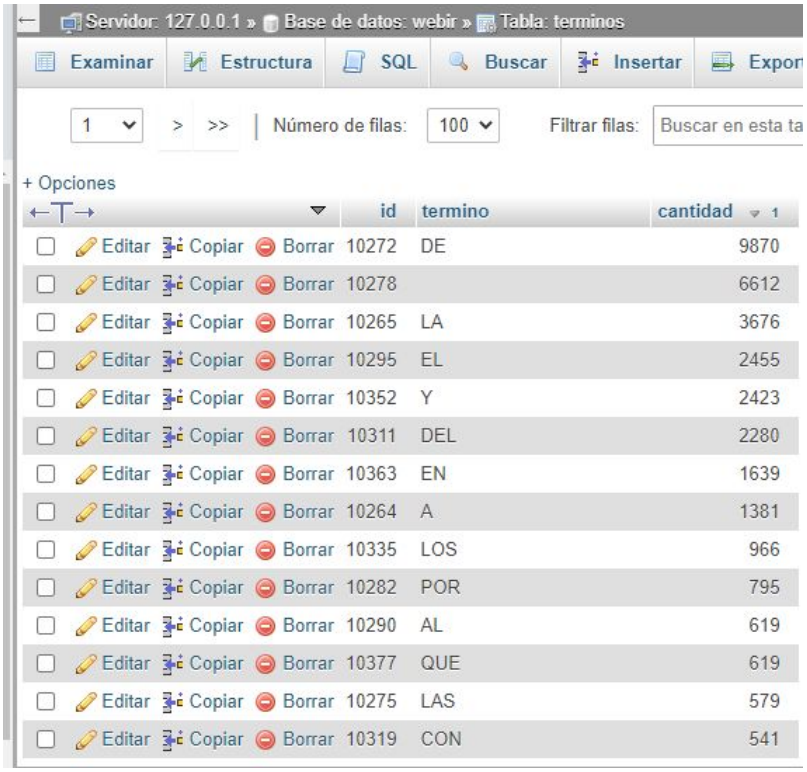
Por cuestiones de tiempo no se logró culminar con la implementación full de un índice invertido, que facilitara otro tipos de análisis sobre la información. Si bien no se cuenta a nivel

de interfaz con la explotación de este recurso, se generó una primera versión de un índice invertido, tokenizando y contabilizando todos los términos encontrados en los nombres de las normas del cuerpo y dentro del primer artículo de cada una de ellas.

Esta tarea en desarrollo nos permitió adelantar algunas conclusiones y visualizar la necesidad de distintos tratamientos posteriores al índice construido.

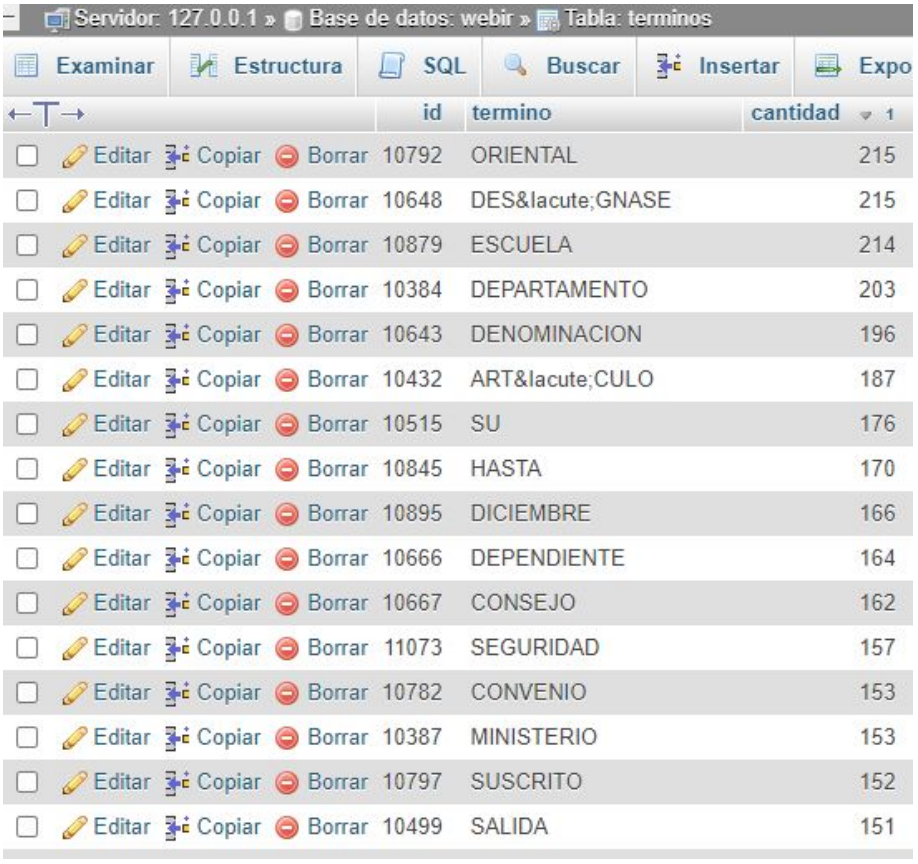
Stop Words

La generación del índice invertido permitió constatar la ocurrencia de stop words como los términos más frecuentes encontrados en el cuerpo analizado. A continuación se detalla un top de términos recuperados.



	id	termino	cantidad
Editar Copiar Borrar	10272	DE	9870
Editar Copiar Borrar	10278		6612
Editar Copiar Borrar	10265	LA	3676
Editar Copiar Borrar	10295	EL	2455
Editar Copiar Borrar	10352	Y	2423
Editar Copiar Borrar	10311	DEL	2280
Editar Copiar Borrar	10363	EN	1639
Editar Copiar Borrar	10264	A	1381
Editar Copiar Borrar	10335	LOS	966
Editar Copiar Borrar	10282	POR	795
Editar Copiar Borrar	10290	AL	619
Editar Copiar Borrar	10377	QUE	619
Editar Copiar Borrar	10275	LAS	579
Editar Copiar Borrar	10319	CON	541

Omitiendo estos términos “sin valor” podemos encontrar con mucha frecuencia términos cómo ORIENTAL, DESÍGNASE, ESCUELA, DEPARTAMENTO, CONSEJO, SEGURIDAD, CONVENIO, etc.



Problemas encontrados

Dentro de la actividad de recuperar la información, encontramos distintas dificultades que suelen ocurrir en tareas de este tipo. A continuación se listan las más significativas y/o las que pudimos atacar:

- Problemas de codificación (tildes, eñe, etc.)
- El retorno json de la API, si bien es muy bien estructurado, no siempre tiene todos los tags, por las propias características de los datos, lo que requirió tratamiento
- Formato de las fecha incompatible con DB requirió conversión
- Formato html para tildes y eñes requirió tratamiento
- Textos en mayúsculas y minúsculas, requirieron normalización para lograr la consulta temática y la construcción del índice invertido
- Los firmantes no vienen en un tag separado, son un string que los contiene a todos. Esto nos obligó a parsear este dato para construir la colección de firmantes de cada norma. Durante el parseo y con el avance en la extracción de los datos, encontramos que dentro del string no siempre aparecían los datos separados de la misma forma:
    - Separadores distintos ';' '-' '.'
    - Saltos de línea que también generaron problemas
- En el índice inverso
    - Palabras especiales, espacios dobles, stop words

Optimizaciones realizadas

Algunas optimizaciones que se realizaron sobre la información recuperada y/o en la consulta de la misma:

- Normalizar a mayúscula para la consulta de términos
- Transformaciones como reemplazo de caracteres: tildes, espacios dobles, etc.

Otras opciones no implementadas

Por cuestiones de tiempo, quedaron como otras optimizaciones a futuro, algunas otras opciones:

- Otras formas de Normalización
- Equivalencias: Por ejemplo a nivel de firmantes, que aparecen de distinta forma en un determinado momento (Presidente => Tabaré Vazquez)
- Conjugaciones
- Familias de palabras
- Stemming
- Lemmatization
- Índices pares de palabras (o frases)
- Índices posicionales

Una herramienta que podría facilitar algunas de estas optimizaciones y que no llegamos a integrar es Elastic Search. Es una opción que en poco tiempo podríamos incorporar a la solución y agregar con bajo esfuerzo, gran potencia en las búsquedas.

Conclusiones y trabajo a futuro

Encontramos que llegar a un enorme conjunto de información no fue difícil, el Estado disponibiliza a través del IMPO y de otras numerosas webs, información valiosa para el estudio del cuerpo de normas existentes. En contrapartida, no hay herramientas que permitan al ciudadano hacer otros tipos de análisis, de carácter más estadístico y cuantitativo, respecto a la producción legislativa.

Logramos recuperar exhaustivamente la normativa disponible, incluso más allá de lo que esperábamos en un momento inicial, aunque en un futuro se podría trabajar mucho más en la precisión de las función de consulta provista, agregando optimizaciones como las mencionadas anteriormente.

Existe un espacio muy interesante de trabajo que podría realizarse en relación a la temática, para generar insumos para que los ciudadanos realicen análisis de uno de los aspectos claves de la democracia.

Un conjunto de tareas que podrían implementarse con los datos ya disponibles, son los siguientes:

- Construir temáticas tabuladas, por ejemplo: A través de relevance feedback, búsquedas usuales, filtrado colaborativo

- Agregar tolerancia a errores de ortografía y otras inconsistencias
- Analizar otras métricas posibles:
    - ¿cantidad de artículos por ley?
    - ¿cantidad de palabras?
    - ¿cantidad de firmantes?
- Agregar otros operadores lógicos a la búsqueda temática (OR, NOT, XOR)
- Calificar resultados por la distancia entre los términos
- Ordenar el retorno por relevancia

Más posibilidades a futuro surgirían de la integración con otras fuentes de datos, que por ejemplo permitan agregar información de las personas firmantes, cómo ser:

- Partido
- Lista
- Sexo
- Raza
- Edad
- Profesión
- Votos a determinados proyectos
- Asistencias a sala

Este tipo de información permitiría generar nuevas dimensiones de análisis, cómo por ejemplo:

- Métricas por partido, lista, etc.
- Participación de la mujer en la producción legislativa
- Participación de minorías étnicas, sexuales, religiosas
- Entender mejor las asociaciones entre firmantes, ¿hay colaboración entre partidos?

Referencias

- https://www.impo.com.uy/
- https://parlamento.gub.uy/